

Generating and Validating Synthetic Unbalanced Panel Data: Evidence from Italian Firm Microdata

Marco Langiulli, Alessandro Moro, Mattia Orzincolo,
and Daniele Piras

Banca d'Italia¹

Lisbon - September 15, 2026

BdP-IFC Workshop on Data Security and Privacy

¹The opinions expressed are those of the authors and do not necessarily reflect the views of the Banca d'Italia.

Motivation and contribution

- Currently, the Bdl's RDC disseminates data through two channels: Remote EXecution (REX) and on-site lab (labBI).
- A virtual lab populated with synthetic data would combine the advantages of both REX (accessibility from anywhere, at any time) and labBI (direct access and processing of data).
- We generate a synthetic version of the Survey of Industrial and Service Firms (INVIND), considering a set of key quantitative and qualitative variables (potential indirect identifiers).
- We extend existing methodologies, originally designed for cross-sectional data, to an unbalanced panel structure.
- We show that measurement errors can be introduced to improve the trade-off between fitting accuracy and disclosure risk.

Data

- Data from the Survey of Industrial and Service Firms (INVIND).
- 13,837 firms observed over the 1993–2024 period.
- Repeated cross-sections, but...
- ...some firms are in multiple waves: unbalanced panel structure.
- We define a threshold θ to separate the cross-sectional and panel components.
- Firms with $T_i \geq \theta$ are classified in the panel component.

INVIND variables used in the analysis.

Description	Symbol
Firm identifier (tax id)	ι_i
Calendar year	t
Geographic area (4 levels)	A_i
Industry sector (11 levels)	S_i
Founding year	F_i
Employment (cur., lag)	$L_{i,t}, L_{i,t-1}$
Hours worked (cur., lag)	$H_{i,t}, H_{i,t-1}$
Turnover (cur., lag, exp.)	$R_{i,t}, R_{i,t-1}, R_{i,t+1}^e$
Export turnover (cur., lag, exp.)	$X_{i,t}, X_{i,t-1}, X_{i,t+1}^e$
Investment (cur., lag, exp.)	$I_{i,t}, I_{i,t-1}, I_{i,t+1}^e$
Price variation (cur., exp.)	$\Delta P_{i,t}, \Delta P_{i,t+1}^e$

- We set $\theta = 10$ years.
- Robustness: $\theta \in \{2, \dots, 20\}$.
- With $\theta = 10$, 31% of firms (65% of observations) are in the panel component.

Sequential modelling: trees

- Modelling of conditional distributions via trees (Reiter, 2005; Kinney et al., 2011; Nowok et al., 2016).
- **Cross-sectional component** ($T_i < \theta$). For each year t , and for each variable j , we estimate the following CART model:

$$Y_{i,t}^{(j)} = f_t^{(j)} \left(Y_{i,t}^{(j-1)}, \dots, Y_{i,t}^{(1)} \right) + \varepsilon_{i,t}^{(j)}. \quad (1)$$

- **Panel component** ($T_i \geq \theta$). For each variable j , we estimate the following CART model:

$$Y_{i,t}^{(j)} = f^{(j)} \left(Y_{i,t}^{(j-1)}, \dots, Y_{i,t}^{(1)}, \iota_i, t \right) + \varepsilon_{i,t}^{(j)}. \quad (2)$$

- synthetic $\tilde{Y}^{(1)}$ is randomly sampled from observed data, $\tilde{Y}^{(2)} = \tilde{f}^{(2)} \left(\tilde{Y}^{(1)} \right) + \tilde{\varepsilon}^{(2)}$, $\tilde{Y}^{(3)} = \tilde{f}^{(3)} \left(\tilde{Y}^{(2)}, \tilde{Y}^{(1)} \right) + \tilde{\varepsilon}^{(3)}$, ...
- Order: time-invariant variables, lags, contemporaneous levels, expectations.

Joint modelling: Gaussian copulas

- Modelling of the joint distribution via copulas (Patki et al., 2016; Gambs et al., 2021; Meyer et al., 2021; Hofert et al., 2026).
- **Cross-sectional component** ($T_i < \theta$). For each year t , area a , and sector s , we model the joint distribution as:

$$F\left(Y_{i,t}^{(1)}, \dots, Y_{i,t}^{(J)}\right) = C_{\Sigma_{a,s,t}}\left(F_1\left(Y_{i,t}^{(1)}\right), \dots, F_J\left(Y_{i,t}^{(J)}\right)\right). \quad (3)$$

- **Panel component** ($T_i \geq \theta$). For each area a , and sector s , we model the joint distribution (including time) as:

$$F\left(Y_{i,t}^{(1)}, \dots, Y_{i,t}^{(J)}, t\right) = C_{\Sigma_{a,s}}\left(F_1\left(Y_{i,t}^{(1)}\right), \dots, F_J\left(Y_{i,t}^{(J)}\right), G(t)\right). \quad (4)$$

- Where F is the joint distribution, F_j and G are the marginal distributions, and $C_{\Sigma}(\cdot)$ is the Gaussian Copula.
- Synthetic values $\tilde{Y}_{i,t}^{(j)}$ are drawn from the estimated joint distribution.

Measurement errors

- In order to increase privacy protection, we add log-normal noise to continuous synthesized variables:

$$\tilde{Y}_{i,t}^{*(j)} = \tilde{Y}_{i,t}^{(j)} \exp(\psi \sigma_{\log(\tilde{Y}^{(j)})} \eta_{i,t}^{(j)}), \quad \eta_{i,t}^{(j)} \sim \mathcal{N}(0, 1), \quad (5)$$

where $\sigma_{\log(\tilde{Y}^{(j)})}$ is the empirical standard deviation of $\log(\tilde{Y}^{(j)})$.

- Normal noise is added to variables which may take positive or negative values:

$$\tilde{Y}_{i,t}^{*(j)} = \tilde{Y}_{i,t}^{(j)} + \psi \sigma_{\tilde{Y}^{(j)}} \eta_{i,t}^{(j)}, \quad \eta_{i,t}^{(j)} \sim \mathcal{N}(0, 1), \quad (6)$$

where $\sigma_{\tilde{Y}}$ is the empirical standard deviation of $\tilde{Y}^{(j)}$.

- $\eta_{i,t}^{(j)}$ is stochastically independent of $\tilde{Y}_{i,t}^{(j)}$.
- $\psi > 0$ is a global tuning parameter calibrated to 0.20.

Distributions

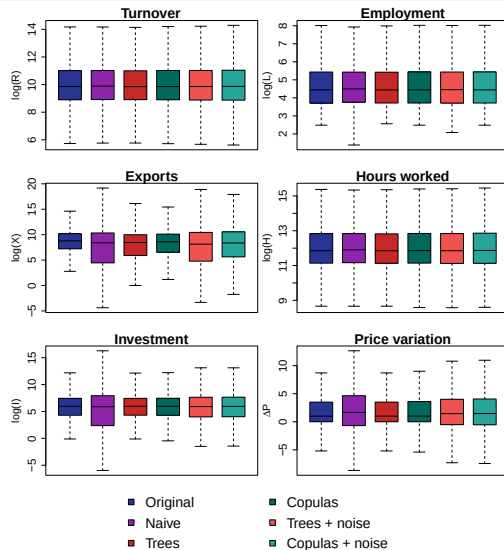


Figure 1: Original vs synthetic univariate distributions.

Random regressions

- We select at random one dependent variable, a subset of regressors, and a set of fixed effects (area, sector, time, firm).
- For each coefficient we compute (Woo and Slavkovic, 2015; Snoke et al., 2018):

$$E = \frac{\hat{\beta}_{\text{syn}} - \hat{\beta}_{\text{orig}}}{\text{sd}(\hat{\beta}_{\text{orig}})}.$$

- We estimate 100 randomly generated specifications, comparing around 3,500 coefficients.

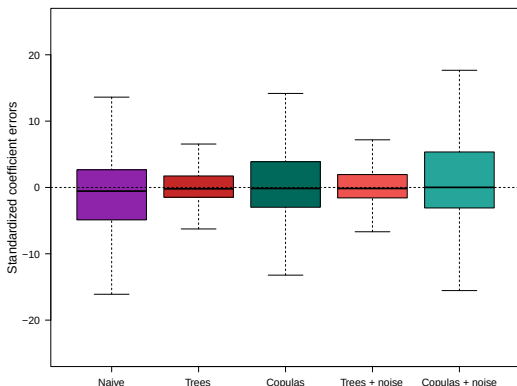


Figure 2: Standardized coefficient errors from random regressions.

Re-identification risk

- The goal is to evaluate disclosure risk by checking how close each synthetic record is to its nearest real record.
- Quasi-identifiers ($\mathbf{X}_{i,t} \subseteq \mathbf{Y}_{i,t}$): year, geographic area, economic sector, employment, turnover, exports, and investment.
- For each synthetic record ($\mathbf{x}_{i,t}^S$), we compute Euclidean distances to its first ($\mathbf{x}_{j,s}^{R1}$) and second ($\mathbf{x}_{l,m}^{R2}$) nearest neighbors in the original data.
- Nearest Neighbor Distance Ratio (NNDR; Lowe, 2004):

$$NNDR(\mathbf{x}_{i,t}^S) = \frac{\|\mathbf{x}_{j,s}^{R1} - \mathbf{x}_{i,t}^S\|}{\|\mathbf{x}_{l,m}^{R2} - \mathbf{x}_{i,t}^S\|} \in (0, 1) \quad (7)$$

- Low NNDR means high disclosure risk: a synthetic record is unusually close to a unique original record.
- Robustness: Gower's distance.

NNDR: results

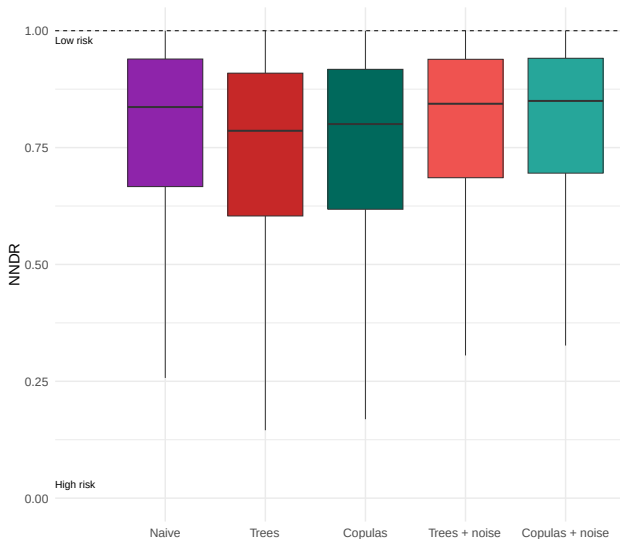


Figure 3: NNDR by method.

Conclusions

- Both sequential and joint modelling methods reproduce the original data well in terms of:
 - distributions;
 - regression coefficients.
- The strong similarity to the original data results in a higher re-identification risk.
- Carefully calibrated noise increases disclosure protection to levels comparable with less informative generators, while preserving key statistical properties.
- High-quality synthetic panels provide an intermediate access layer for RDCs, enabling exploration, empirical design, and code debugging before running analyses on confidential microdata.